Version with Markings to Show Changes Made

1. A method for managing caches in a system with multiple caches that may contain different copies of a data item, comprising the steps of:

modifying the data item in a first node of said multiple caches to create a modified data item;

sending the modified data item from said first node to a second node of said multiple caches without durably storing the modified data item from said first node to persistent storage;

after said modified data item has been sent from said first node to said second node, said first node sending a request to a master of said data item for writing said data item to persistent storage; and

in response to said request, said master coordinating with said multiple caches to cause said data item to be written to persistent storage.

2. The method of Claim 1 wherein:

the method includes the step of maintaining, within an ordered series of bins, entries for past-image versions of data items;

each bin in said ordered series corresponds to a time range;

a particular bin corresponds to the time range that covers the time at which the data item is modified in said first node; and

the step of sending a request is performed by sending a request for writing a particular bin of said ordered series of bins to persistent storage.

3. The method of Claim 2 wherein the step of said master coordinating with said multiple caches to cause said data item to be written to persistent storage includes said master causing said multiple caches to write data items to persistent storage to

cover all past image versions of data items that were modified during the time range of said particular bin.

4. The method of Claim 3 further comprising the step of emptying said particular bin after said multiple caches write data items to persistent storage to cover all past image versions of data items that were modified during the time range of said particular bin.

5. The method of Claim 4 wherein the step of emptying said particular bin includes the steps of:

discarding entries within said particular bin that are associated with past images that have last-dirtied times within the time range of said particular bin; and

moving to one or more other bins the entries within said particular bin that are associated with past images that have last-dirtied times later than the time range of said particular bin.

6. The method of Claim 1 wherein the step of sending a request to a master is performed by sending the request to a global lock manager.

7. The method of Claim 1 wherein the step of sending a request to a master is performed by sending the request to a lock manager that is one of a plurality of lock managers within a distributed lock management system.

8. The method of Claim 1 further comprising the step of sending from the master, to interested nodes, write-notification messages indicating that said data item has been written to persistent storage, in response to said data item being written to persistent storage.

9. The method of Claim 8 wherein the step of sending write-notification messages includes the master sending to at least one interested node a single message that notifies said at least one interested node that a plurality of data items have been written to persistent storage.

10. The method of Claim 1 wherein the step of said first node sending a request to a master of said data item for writing said data item to persistent storage includes the first node sending to said master a single message that requests writing a plurality of data items to persistent storage, wherein said plurality of data items includes said data item.

11. The method of Claim 10 wherein the step of sending a single message includes sending a message that identifies a bin to request that all data items that belong to the bin be written to persistent storage.

12. The method of Claim 11 wherein the bin is associated with a range of time and includes data items that were first dirtied by the first node during said range of time and that were subsequently transferred to other nodes without first being written to persistent storage.

13. The method of Claim 8 wherein the step of sending from the master to interested nodes write-notification messages includes the steps of:

immediately sending write-notification messages to a first set of interested nodes, where said first set of interested nodes includes the interested nodes that have requested said data item to be written to persistent storage; and

delaying the sending of write-notification messages to a second set of nodes, where said second set of nodes includes interested nodes that do not belong to said first set of interested nodes.

14. The method of Claim 8 wherein the step of sending from the master to interested nodes includes delaying the sending of write-notification messages to at least one interested node.

15. The method of Claim 14 wherein a write-notification message is sent to the at least one interested node in response to a lock request made by said at least one interested node.

16. The method of Claim 14 wherein a write-notification message is sent to the at least one interested node in response to the at least one interested node requesting that said data item be written to persistent storage.

17. The method of Claim 14 wherein a write-notification message is sent to the at least one interested node within a ping request that the master sends to the at least one interested node for the at least one interested node to transfer another data item to another node.

18. The method of Claim 1 wherein the step of coordinating includes the steps of:
determining whether a version of said data item, that is at least as recent as said modified version, has already been written to persistent storage; and
if a version of said data item that is at least as recent as said modified version has already been written to persistent storage, then sending a write-notification message from said master to notify said first node that a version of said data

item that is at least as recent as said modified version has already been written to persistent storage.

19. The method of Claim 18 wherein the step of coordinating includes, if a version of said data item that is at least as recent as said modified version has not already been written to persistent storage, then sending a write-perform message from said master to grant permission for said modified version to be written to persistent storage.

20. The method of Claim 1 wherein the step of coordinating includes the steps of:

selecting a particular node of said multiple caches that has a particular version of said data item, wherein said particular version is at least as recent as the modified data item in said first node; and

causing said particular version of said data item to be written from said particular node to persistent storage.

21. The method of Claim 20 wherein the step of selecting a particular node includes selecting the node, of said multiple caches, that has a most recently modified version of said data item.

22. The method of Claim 20 further comprising the step of the master informing the first node that said data item has been written to persistent storage in response to the master receiving confirmation that said particular version of said data item has been written to persistent storage.

23. The method of Claim 20 further comprising the step of the master informing a set of caches that said data item has been written to persistent storage in response to the master receiving confirmation that said particular version of said data item has been

written to persistent storage, wherein said set of caches includes caches, other than said particular node, that contain modified versions of said data item that are not more recent than said particular version.

24. A method for managing caches in a system with multiple caches that may contain different copies of a data item, comprising the steps of:

modifying the data item in a first cache to create a modified data item;

in response to writing the modified data item to persistent storage, performing the steps of:

a node associated with the first cache determining whether any other cache in said multiple caches had created a dirty version of said data item; and

if any other cache in said multiple caches had created a dirty version of said data item, then the node associated with the first cache informing a master of said data item that said modified data item has been written to persistent storage; and

if no other cache in said multiple caches had created a dirty version of said data item, then the step of writing the modified data item is performed without informing said master that said modified data item has been written to persistent storage.

25. The method of Claim 24 wherein the step of determining whether any other cache in said multiple caches had created a dirty version of said data item includes inspecting a global dirty flag associated with said data item.

26. The method of Claim 25 wherein:

prior to modifying the data item in said first cache, the data item had been modified in a second cache;

the data item is not persistently stored between being modified in said second cache and being modified in said first cache; and

a node associated with the second cache causes said global dirty flag to be set to indicate that said data item is globally dirty.

27. The method of Claim 26 wherein the node associated with the second cache causes said global dirty flag to be set in response to transferring said data item from said second cache to another cache of said multiple caches.

28. A method for managing caches in a system with multiple caches that may contain different copies of a data item, comprising the steps of:

modifying the data item in a first cache to create a modified data item;

when a node associated with the first node desires to write said modified data item to persistent storage, performing the steps of:

if then node associated with the first cache does not currently have ownership rights to said data item, then the node associated with the first cache sending a request to a master of said data item for said data item to be written to persistent storage; and

if said node associated with said first cache currently has ownership rights to said data item, then the first node writing said data item to persistent storage without sending a request to said master for said data item to be written to persistent storage.

~~28~~29. The method of Claim ~~27~~28 further comprising the step of designating holders of exclusive locks in data items to be owners of said data items.

~~29~~30. A method for managing a data item, the method comprising the steps of:

when a node that has an exclusive lock on a data item desires to write the data item to persistent storage, performing the steps of

determining whether a mode associated with the data item is local or global;

if the mode associated with the data item is local, then the node writing the data item to persistent storage without communicating with a master of said data item; and

if the mode associated with the data item is global, then the node sending a message to the master of the data item to request writing of said data item to persistent storage.

~~30~~31. The method of Claim ~~29~~30 wherein:

the mode associated with the data item is global; and

the method further includes:

the node receiving permission from the master to write the data item to persistent storage; and

after writing the data item to persistent storage, changing the mode from global to local.

~~31~~32. The method of Claim ~~29~~30 wherein:

the mode associated with the data item is local; and

before the node has completed writing of the data item to persistent storage, the node transfers the exclusive lock on the data item to another node.

~~32~~33. The method of Claim ~~31~~32 wherein the node changes the mode from local to global prior to transferring the exclusive lock on the data item to another node.

~~33~~34. The method of Claim ~~31~~32 wherein the node informs the master when the node has completed writing the data item to persistent storage.

3435. The method of Claim 3132 wherein the master informs the other node that the node has completed writing the data item to persistent storage in response to the node informing the master that the node has completed writing the data item to persistent storage.

3536. The method of Claim 3233 wherein the other node changes the mode from global to local in response to a message from the master after the node has informed the master that the node has completed writing the data item to persistent storage.

3637. A method for managing a data item, the method comprising the steps of:

when a data item is transferred from one node to another node, performing the steps of

if the data item has been dirtied by the node and a mode associated with the data item is local, then changing the mode from local to global prior to sending the data item to another node;

if the data item has not been dirtied by the node and the mode associated with the data item is local, then sending the data item to the other node without changing the mode;

allowing the other node to write the data item to persistent storage without requesting permission if the mode is local; and

requiring the other node to obtain permission to write the data item to persistent storage if the mode is global.

3738. The method of Claim 3637 wherein the node transfers the data item to the other node prior to completion of the node writing the data item to persistent storage.

3839. The method of Claim ~~3738~~ wherein, after completion of the node writing the data item to persistent storage, the node sends a message to a master of the data item to indicate that the data item has been written to persistent storage.

3940. The method of Claim ~~3839~~ wherein:

the other node receives the data item in global mode; and

the other node sends a request to the master of the node for permission to write the data item; and

the master responds to said request by informing said other node to change said mode from global to local.

4041. A method for managing versions of a data item, the method comprising the steps of:

when a dirty version of a data item is transferred from a first node to a second node while a being-written version of the data item is being written to persistent storage, performing the steps of:

communicating version information about the being-written version to the second node; and

based on the version information, the second node preventing any version of the data item that belongs to a first set of versions from being merged with any version of the data item that belongs to a second set of versions;

wherein the first set of versions includes all versions of the data item within the second node that are at least as old as the being-written version; and

wherein the second set of versions includes versions of the data item within the second node that are newer than the being-written version.

4142. The method of Claim 4041 wherein the step of communicating is performed by a master assigned to said data item.

4243. The method of Claim 4041 wherein:
the second node includes a plurality of versions in said first set; and
the second node merges said plurality of versions.

4344. The method of Claim 4041 further comprising the steps of:
informing the second node when the being-written version has been successfully written to persistent storage; and
after the second node has been informed that the being-written version has been successfully written to persistent storage, allowing said second node to discard all versions in said first set of versions.

4445. The method of Claim 4243 further comprising the steps of:
informing the second node when the being-written version has been successfully written to persistent storage; and
after the second node has been informed that the being-written version has been successfully written to persistent storage, allowing said second node to discard a merged version created by merging said plurality of versions.

4546. A method for managing past images of a data item, the method comprising the steps of:
estimating a likelihood that a first past version of a data item will soon be written to persistent storage or covered by a write to persistent storage;

if the estimated likelihood is exceeds a particular threshold, then storing a second past version of the data item separate from the first past version of the data item; and

if the estimated likelihood falls below a particular threshold, then merging the second past version of the data item with the first past version of the data item.

4647. The method of Claim 4546 wherein the step of estimating is based on a comparison between a time associated with the first past version of the data item and a time associated with a recent entry in a redo log file.

4748. The method of Claim 4546 wherein the step of estimating is based on a comparison between a time associated with the first past version of the data item and a time associated with an entry at the head of a checkpoint queue.

4849. A computer-readable medium carrying instructions for managing caches in a system with multiple caches that may contain different copies of a data item, the instructions comprising instructions for performing the steps of:

modifying the data item in a first node of said multiple caches to create a modified data item;

sending the modified data item from said first node to a second node of said multiple caches without durably storing the modified data item from said first node to persistent storage;

after said modified data item has been sent from said first node to said second node, said first node sending a request to a master of said data item for writing said data item to persistent storage; and

in response to said request, said master coordinating with said multiple caches to cause said data item to be written to persistent storage.

4950. The computer-readable medium of Claim 4849 wherein:

the computer-readable medium includes instructions for performing the step of maintaining, within an ordered series of bins, entries for past-image versions of data items;

each bin in said ordered series corresponds to a time range;

a particular bin corresponds to the time range that covers the time at which the data item is modified in said first node; and

the step of sending a request is performed by sending a request for writing a particular bin of said ordered series of bins to persistent storage.

5051. The computer-readable medium of Claim 4950 wherein the step of said master coordinating with said multiple caches to cause said data item to be written to persistent storage includes said master causing said multiple caches to write data items to persistent storage to cover all past image versions of data items that were modified during the time range of said particular bin.

5152. The computer-readable medium of Claim 5051 further comprising instructions for performing the step of emptying said particular bin after said multiple caches write data items to persistent storage to cover all past image versions of data items that were modified during the time range of said particular bin.

5253. The computer-readable medium of Claim 5152 wherein the step of emptying said particular bin includes the steps of:

discarding entries within said particular bin that are associated with past images that have last-dirtied times within the time range of said particular bin; and

moving to one or more other bins the entries within said particular bin that are associated with past images that have last-dirtied times later than the time range of said particular bin.

~~53~~54. The computer-readable medium of Claim ~~48~~49 wherein the step of sending a request to a master is performed by sending the request to a global lock manager.

~~54~~55. The computer-readable medium of Claim ~~48~~49 wherein the step of sending a request to a master is performed by sending the request to a lock manager that is one of a plurality of lock managers within a distributed lock management system.

~~55~~56. The computer-readable medium of Claim ~~48~~49 further comprising instructions for performing the step of sending from the master, to interested nodes, write-notification messages indicating that said data item has been written to persistent storage, in response to said data item being written to persistent storage.

~~56~~57. The computer-readable medium of Claim ~~55~~56 wherein the step of sending write-notification messages includes the master sending to at least one interested node a single message that notifies said at least one interested node that a plurality of data items have been written to persistent storage.

~~57~~58. The computer-readable medium of Claim ~~48~~49 wherein the step of said first node sending a request to a master of said data item for writing said data item to persistent storage includes the first node sending to said master a single message that requests writing a plurality of data items to persistent storage, wherein said plurality of data items includes said data item.

5859. The computer-readable medium of Claim 5758 wherein the step of sending a single message includes sending a message that identifies a bin to request that all data items that belong to the bin be written to persistent storage.

5960. The computer-readable medium of Claim 5859 wherein the bin is associated with a range of time and includes data items that were first dirtied by the first node during said range of time and that were subsequently transferred to other nodes without first being written to persistent storage.

6061. The computer-readable medium of Claim 5556 wherein the step of sending from the master to interested nodes write-notification messages includes the steps of:

immediately sending write-notification messages to a first set of interested nodes, where said first set of interested nodes includes the interested nodes that have requested said data item to be written to persistent storage; and

delaying the sending of write-notification messages to a second set of nodes, where said second set of nodes includes interested nodes that do not belong to said first set of interested nodes.

6162. The computer-readable medium of Claim 5556 wherein the step of sending from the master to interested nodes includes delaying the sending of write-notification messages to at least one interested node.

6263. The computer-readable medium of Claim 6162 wherein a write-notification message is sent to the at least one interested node in response to a lock request made by said at least one interested node.

6364. The computer-readable medium of Claim 6162 wherein a write-notification message is sent to the at least one interested node in response to the at least one interested node requesting that said data item be written to persistent storage.

6465. The computer-readable medium of Claim 6162 wherein a write-notification message is sent to the at least one interested node within a ping request that the master sends to the at least one interested node for the at least one interested node to transfer another data item to another node.

6566. The computer-readable medium of Claim 4849 wherein the step of coordinating includes the steps of:

determining whether a version of said data item, that is at least as recent as said modified version, has already been written to persistent storage; and

if a version of said data item that is at least as recent as said modified version has already been written to persistent storage, then sending a write-notification message from said master to notify said first node that a version of said data item that is at least as recent as said modified version has already been written to persistent storage.

6667. The computer-readable medium of Claim 6566 wherein the step of coordinating includes, if a version of said data item that is at least as recent as said modified version has not already been written to persistent storage, then sending a write-perform message from said master to grant permission for said modified version to be written to persistent storage.

6768. The computer-readable medium of Claim 4849 wherein the step of coordinating includes the steps of:

selecting a particular node of said multiple caches that has a particular version of said data item, wherein said particular version is at least as recent as the modified data item in said first node; and

causing said particular version of said data item to be written from said particular node to persistent storage.

~~68~~69. The computer-readable medium of Claim ~~67~~68 wherein the step of selecting a particular node includes selecting the node, of said multiple caches, that has a most recently modified version of said data item.

~~69~~70. The computer-readable medium of Claim ~~67~~68 further comprising instructions for performing the step of the master informing the first node that said data item has been written to persistent storage in response to the master receiving confirmation that said particular version of said data item has been written to persistent storage.

~~70~~71. The computer-readable medium of Claim ~~67~~68 further comprising instructions for performing the step of the master informing a set of caches that said data item has been written to persistent storage in response to the master receiving confirmation that said particular version of said data item has been written to persistent storage, wherein said set of caches includes caches, other than said particular node, that contain modified versions of said data item that are not more recent than said particular version.

~~71~~72. A computer-readable medium carrying instructions for managing caches in a system with multiple caches that may contain different copies of a data item, the instructions comprising instructions for performing the steps of:

modifying the data item in a first cache to create a modified data item;

in response to writing the modified data item to persistent storage, performing the steps of:

a node associated with the first cache determining whether any other cache in said multiple caches had created a dirty version of said data item; and

if any other cache in said multiple caches had created a dirty version of said data item, then the node associated with the first cache informing a master of said data item that said modified data item has been written to persistent storage; and

if no other cache in said multiple caches had created a dirty version of said data item, then the step of writing the modified data item is performed without informing said master that said modified data item has been written to persistent storage.

~~72~~73. The computer-readable medium of Claim ~~71~~72 wherein the step of determining whether any other cache in said multiple caches had created a dirty version of said data item includes inspecting a global dirty flag associated with said data item.

~~73~~74. The computer-readable medium of Claim ~~72~~73 wherein:

prior to modifying the data item in said first cache, the data item had been modified in a second cache;

the data item is not persistently stored between being modified in said second cache and being modified in said first cache; and

a node associated with the second cache causes said global dirty flag to be set to indicate that said data item is globally dirty.

~~74~~75. The computer-readable medium of Claim ~~73~~74 wherein the node associated with the second cache causes said global dirty flag to be set in response to transferring said data item from said second cache to another cache of said multiple caches.

~~75~~76. A computer-readable medium carrying instructions for managing caches in a system with multiple caches that may contain different copies of a data item, the instructions comprising instructions for performing the steps of:

modifying the data item in a first cache to create a modified data item;

when a node associated with the first node desires to write said modified data item to persistent storage, performing the steps of:

if then node associated with the first cache does not currently have ownership rights to said data item, then the node associated with the first cache sending a request to a master of said data item for said data item to be written to persistent storage; and

if said node associated with said first cache currently has ownership rights to said data item, then the first node writing said data item to persistent storage without sending a request to said master for said data item to be written to persistent storage.

~~75~~77. The computer-readable medium of Claim ~~74~~76 further comprising instructions for performing the step of designating holders of exclusive locks in data items to be owners of said data items.

~~76~~78. A computer-readable medium carrying instructions for managing a data item, the instructions comprising instructions for performing the steps of:

when a node that has an exclusive lock on a data item desires to write the data item to persistent storage, performing the steps of

determining whether a mode associated with the data item is local or global;

if the mode associated with the data item is local, then the node writing the data item to persistent storage without communicating with a master of said data item; and

if the mode associated with the data item is global, then the node sending a message to the master of the data item to request writing of said data item to persistent storage.

~~77~~79. The computer-readable medium of Claim ~~76~~78 wherein:

the mode associated with the data item is global; and

the computer-readable medium further includes instructions for:

the node receiving permission from the master to write the data item to persistent storage; and

after writing the data item to persistent storage, changing the mode from global to local.

~~78~~80. The computer-readable medium of Claim ~~76~~78 wherein:

the mode associated with the data item is local; and

before the node has completed writing of the data item to persistent storage, the node transfers the exclusive lock on the data item to another node.

~~79~~81. The computer-readable medium of Claim ~~78~~80 wherein the node changes the mode from local to global prior to transferring the exclusive lock on the data item to another node.

~~80~~82. The computer-readable medium of Claim ~~78~~80 wherein the node informs the master when the node has completed writing the data item to persistent storage.

8183. The computer-readable medium of Claim 7880 wherein the master informs the other node that the node has completed writing the data item to persistent storage in response to the node informing the master that the node has completed writing the data item to persistent storage.

8284. The computer-readable medium of Claim 7981 wherein the other node changes the mode from global to local in response to a message from the master after the node has informed the master that the node has completed writing the data item to persistent storage.

8385. A computer-readable medium carrying instructions for managing a data item, the instructions comprising instructions for performing the steps of:

when a data item is transferred from one node to another node, performing the steps of

if the data item has been dirtied by the node and a mode associated with the data item is local, then changing the mode from local to global prior to sending the data item to another node;

if the data item has not been dirtied by the node and the mode associated with the data item is local, then sending the data item to the other node without changing the mode;

allowing the other node to write the data item to persistent storage without requesting permission if the mode is local; and

requiring the other node to obtain permission to write the data item to persistent storage if the mode is global.

8486. The computer-readable medium of Claim 8385 wherein the node transfers the data item to the other node prior to completion of the node writing the data item to persistent storage.

8587. The computer-readable medium of Claim 8486 wherein, after completion of the node writing the data item to persistent storage, the node sends a message to a master of the data item to indicate that the data item has been written to persistent storage.

8688. The computer-readable medium of Claim 8587 wherein:
the other node receives the data item in global mode; and
the other node sends a request to the master of the node for permission to write the data item; and
the master responds to said request by informing said other node to change said mode from global to local.

8789. A computer-readable medium carrying instructions for managing versions of a data item, the instructions comprising instructions for performing the steps of:
when a dirty version of a data item is transferred from a first node to a second node while a being-written version of the data item is being written to persistent storage, performing the steps of:
communicating version information about the being-written version to the second node; and
based on the version information, the second node preventing any version of the data item that belongs to a first set of versions from being merged with any version of the data item that belongs to a second set of versions;

wherein the first set of versions includes all versions of the data item within the second node that are at least as old as the being-written version; and

wherein the second set of versions includes versions of the data item within the second node that are newer than the being-written version.

8890. The computer-readable medium of Claim 8789 wherein the step of communicating is performed by a master assigned to said data item.

8991. The computer-readable medium of Claim 8789 wherein:

the second node includes a plurality of versions in said first set; and

the second node merges said plurality of versions.

9092. The computer-readable medium of Claim 8789 further comprising instructions for performing the steps of:

informing the second node when the being-written version has been successfully written to persistent storage; and

after the second node has been informed that the being-written version has been successfully written to persistent storage, allowing said second node to discard all versions in said first set of versions.

9193. The computer-readable medium of Claim 8991 further comprising instructions for performing the steps of:

informing the second node when the being-written version has been successfully written to persistent storage; and

after the second node has been informed that the being-written version has been successfully written to persistent storage, allowing said second node to discard a merged version created by merging said plurality of versions.

9294. A computer-readable medium carrying instructions for managing past images of a data item, the instructions comprising instructions for performing the steps of:

estimating a likelihood that a first past version of a data item will soon be written to persistent storage or covered by a write to persistent storage;

if the estimated likelihood is exceeds a particular threshold, then storing a second past version of the data item separate from the first past version of the data item; and

if the estimated likelihood falls below a particular threshold, then merging the second past version of the data item with the first past version of the data item.

9395. The computer-readable medium of Claim 9294 wherein the step of estimating is based on a comparison between a time associated with the first past version of the data item and a time associated with a recent entry in a redo log file.

9496. The computer-readable medium of Claim 9294 wherein the step of estimating is based on a comparison between a time associated with the first past version of the data item and a time associated with an entry at the head of a checkpoint queue.